

# EXPERIMENTS ON PARALLEL TRAINING OF DEEP NEURAL NETWORK USING MODEL AVERAGING

Hang Su<sup>1,2</sup>, Haoyu Chen<sup>1</sup>, Haihua Xu<sup>3</sup>

<sup>1</sup> International Computer Science Institute, Berkeley, California, US

<sup>2</sup> Dept. of Electrical Engineering & Computer Science, University of California, Berkeley, CA, USA

<sup>3</sup> Nanyang Technological University, Singapore

{suhang3240@gmail.com}

## ABSTRACT

In this work we apply model averaging to parallel training of deep neural network (DNN). Parallelization is done in a model averaging manner. Data is partitioned and distributed to different nodes for local model updates, and model averaging across nodes is done every few minibatches.

We use multiple GPUs for data parallelization, and Message Passing Interface (MPI) for communication between nodes, which allows us to perform model averaging frequently without losing much time on communication. We investigate the effectiveness of Natural Gradient Stochastic Gradient Descent (NG-SGD) and Restricted Boltzmann Machine (RBM) pretraining for parallel training in model-averaging framework, and explore the best setups in term of different learning rate schedules, averaging frequencies and minibatch sizes. It is shown that NG-SGD and RBM pretraining benefits parameter-averaging based model training. On the 300h Switchboard dataset, a 9.3 times speedup is achieved using 16 GPUs and 17 times speedup using 32 GPUs with limited decoding accuracy loss.<sup>1</sup>

**Index Terms**— Parallel training, model averaging, deep neural network, natural gradient

## 1. INTRODUCTION

Deep Neural Networks (DNN) has shown its effectiveness in several machine learning tasks, especially in speech recognition. The large model size and massive training examples make DNN a powerful model for classification. However, these two factors also slow down the training procedure.

Parallelization of DNN training has been a popular topic since the revival of neural networks. Several different strategies have been proposed to tackle this problem. Multiple thread CPU parallelization and single GPU implementation

are compared in [1, 2], and it is shown that single GPU could beat multi-threaded CPU implementation by a factor of 2.

Optimality for parallelization of DNN training was analyzed in [3], and based on the analysis, a gradient quantization approach (1-bit SGD) was proposed to minimize communication cost [4]. It shows that 1 bit quantization can effectively reduce data exchange in an MPI framework, and a 10 times speed-up is achieved using 40 GPUs.

DistBelief proposed in [5] reports that 8 CPU machines train 2.2 times faster than a single GPU machine on a moderately sized speech model. Asynchronous SGD using multiple GPUs achieved a 3.2x speed-up on 4 GPUs [6].

A pipeline training approach was proposed in [7] and a 3.3x speedup was achieved using 4 GPUs, but this method does not scale beyond number of layers in the neural network.

A speedup of 6x to 14x was achieved using 16 GPUs on training convolutional neural networks [8]. In this approach, each GPU is responsible for a partition of the neural network. This approach is more useful for image classification where local structure of the neural network could be exploited. For a fully connected speech model, a model partition approach may not be able to contribute as much.

Distributed model averaging is used in [9, 10], and a further improvement is done using NG-SGD [11]. In this approach, separate models are trained on multiple nodes using different partitions of data, and model parameters are averaged after each epoch. It is shown that NG-SGD can effectively improve convergence and ensure a better model trained using the model averaging framework.

Our approach is mainly based on the NG-SGD with model averaging. We utilize multiple GPUs in neural networks training via MPI, which allows us to perform model averaging more frequently and efficiently. Unlike the other approach [4], we do not use a warm-up phase where only single thread is used for model update. (Admittedly, this might lead to further improvement). In this work, we conduct a lot of experiments and compare different setups in model averaging framework.

In Section 2, we introduce related works on NG-SGD.

<sup>1</sup>This work is not submitted to peer-review conferences because the authors think it needs more investigation. The authors are in lack of resources to perform further exploration. However, we welcome any comments and suggestions.

Section 3 describe the model averaging approach and some intuition on the analysis. Section 4 records experimental results on different setups and Section 5 concludes.

## 2. RELATIONSHIP TO PRIOR WORKS

To avoid confusion, we should mention that Kaldi[12] contains two neural network recipes. The first implementation<sup>2</sup> is described in [13] which supports Restricted Boltzmann Machine pretraining [14] and sequence-discriminative training [15]. It uses single GPU for SGD training. The second implementation<sup>3</sup> [9] was originally designed to support parallel training on multiple CPUs. Now it also supports multiple GPUs for training using model averaging. By default, it uses layer-wise discriminative pretraining.

Our work extends the first implementation so that it can utilize multiple GPUs using model averaging. We use MPI in implementation, so file I/O is avoided during model averaging. This allows us to perform model averaging much more frequently.

## 3. DATA PARALLELIZATION AND MODEL AVERAGING

SGD is a popular method for DNN training. Even though neural network training objectives are usually non-convex, mini-batch SGD has been shown to be effective for optimizing the objective[16]. Roughly speaking, a bigger minibatch size gives a better estimate of the gradient, resulting in a better the converge rate. Thus, a straight forward idea for parallelization would be distributing the gradient computation to different computing nodes. In each step, gradients of mini-batches on different nodes are reduced to a single node, averaged and then used to update models in each node. This method, i.e. gradient averaging, can compute the gradient accurately, but it requires heavy communication between nodes. Also, it is shown that increasing minibatch size does not always benefit model training[16], especially in early stage of model training.

On the other hand, if we choose to average the parameters rather than gradients, it is not necessary to exchange data that often. Currently, there is no straight forward theory that guarantees convergence, but we would like to explore a bit why this strategy should work, just as we observe in the experiments.

First, in the extreme case where model parameters are averaged after each weight update, model averaging is equivalent to gradient averaging. Furthermore, if model averaging is done every  $n$  minibatch based weight update, model update

formula could be written as

$$\begin{aligned}\theta_{t+n} &= \theta_t + \sum_{i=0}^{n-1} \alpha g_{t+i} \\ &= \theta_t + \sum_{i=0}^{n-1} \alpha \frac{\partial}{\partial \theta} F(x; \theta_{t+i})\end{aligned}\quad (1)$$

$$\overline{\theta}_{t+n} = \overline{\theta}_t + \sum_{i=0}^{n-1} \alpha \frac{\partial}{\partial \theta} \overline{F}(x; \theta_{t+i}) \quad (2)$$

where  $\theta$  is the model parameter and  $\alpha$  is learning rate. If changes in model parameter  $\theta$  is limited within  $n$  updates, this approach could be seen as an approximation to gradient averaging.

Second, it is shown that model averaging for convex models is guaranteed to converge [17, 18]. It is suggested that unsupervised pretraining guides the learning towards basins of attraction of minima that support better generalization from the training data set; [19].

Fig 1 is an example of all-reduce with 4 nodes. This operation could be easily implemented by MPI\_Allreduce.

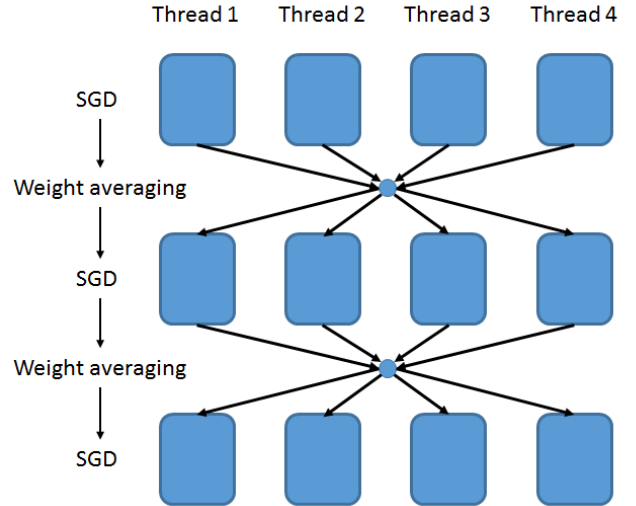


Fig. 1. All-reduce network

## 4. NATURAL GRADIENT FOR MODEL UPDATE

This section introduces the idea proposed in [11].

In stochastic gradient descent (SGD), the learning rate is often assumed to be a scalar  $\alpha_t$  that may change over time, the update formula for model parameters  $\theta_t$  is

$$\theta_{t+1} = \theta_t + \alpha_t g_t \quad (3)$$

where  $g_t$  is the gradient.

<sup>2</sup>Location in code: src/{nnet,nnetbin}

<sup>3</sup>Location in code: src/nnet2,nnet2bin

However, according to Natural Gradient idea [20, 21], it is possible to replace the scalar with a symmetric positive definite matrix  $E_t$ , which is the inverse of the Fisher information matrix.

$$\theta_{t+1} = \theta_t + \alpha_t E_t g_t \quad (4)$$

Suppose  $x$  is the variable we are modeling, and  $f(x; \theta)$  is the probability or likelihood of  $x$  given parameters  $\theta$ , then the Fisher information matrix  $I(\theta)$  is defined as

$$E \left[ \left( \frac{\partial}{\partial \theta} \log f(x; \theta) \right) \left( \frac{\partial}{\partial \theta} \log f(x; \theta) \right)^T \right] \quad (5)$$

For large scale speech recognition, it is impossible to estimate Fisher information matrix and perform inversion, so it is necessary to approximate the inverse Fisher information matrix directly. Details about the theory and implementation of NG-SGD could be found in [11].

## 5. EXPERIMENTAL RESULTS

### 5.1. Setup

In this work, we report speech recognition results on the 300 hour Switchboard conversational telephone speech task (Switchboard-1 Release 2). We use MSU-ISIP release of the Switchboard segmentations and transcriptions (date 11/26/02), together with the Mississippi State transcripts2 and the 30Kword lexicon released with those transcripts. The lexicon contains pronunciations for all words and word fragments in the training data. We use the Hub5 00 data for evaluation. Specifically, we use the the development set and Hub5 01 (LDC2002S13) data as a separate test set.

The Kaldi toolkit[12] is used for speech recognition framework. Standard 13-dim PLP feature, together with 3-dim Kaldi pitch feature, is extracted and used for maximum likelihood GMM model training. Features are then transformed using LDA+MLLT before SAT training. After GMM training is done, a tanh-neuron DNN-HMM hybrid system is trained using the the 40-dimension transformed fMLLR (also known as CMLLR [22]) feature as input and GMM-aligned senones as targets. fMLLR is estimated in an EM fashion for both training data and test data. A trigram language model (LM) is trained on 3M words of the training transcripts only.

Work in this paper is built on top of the Kaldi nnet1 setup and the NG-SGD method introduced in nnet2 setup. Details of DNN training follows Section 2.2 in [13]. In this work, we use 6 hidden layers, where each hidden layer has 2048 neurons with sigmoids. Input layer is 440 dimension (i.e. the context of 11 fMLLR frames), and output layer is 8806 dimension. Mini-batch SGD is used for backpropagation and the minibatch is set to 1024 for all the experiments. By default, DNNs are initialized with stacked restricted Boltzmann machines (RBMs) that are pretrained in a greedy layerwise fashion [14]. Comparison between random initialization and

RBM-initialization in model averaging framework is reported in Section 5.3.

The server hardware used in this work is Stampede (TACC) (URL: <https://portal.xsede.org/tacc-stampede>). It is a Dell Linux cluster provided as an Extreme Science Engineering Discovery Environment (XSEDE) digital service by the Texas Advanced Computing Center (TACC). Stampede is configured with 6,400 Dell DCS Zeus compute nodes, the majority of which are configured with two 2.7 GHz E5-2680 Intel Xeon (Sandy Bridge) processors and one Intel Xeon Phi SE10P coprocessor. 128 of the nodes are augmented with an NVIDIA K20 GPU and 8 GB of GDDR5 memory each, which we use for neural network training in this work. Stampede nodes run Linux 2.6.32 with batch services managed by the Simple Linux Utility for Resource Management (SLURM).

### 5.2. Switchboard Results

Fig. 2 shows the scaling factor and speedup plot for model averaging experiments. As is shown in the graph, a speedup of 17 could be achieved when 32 GPUs are used. Table 1

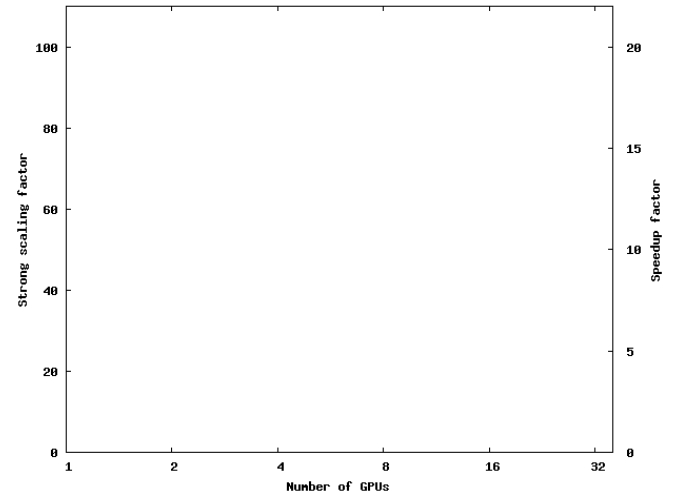


Fig. 2. Scaling factor and speedup factor v.s. number of gpus

shows the main decoding results for DNNs trained using different number of GPUs. In general, decoding results of DNNs trained model averaging degrades 0.3~0.4 WER, depending on the number of GPUs used.

### 5.3. Initialization Matters

Table 2 compares random initialization with Restricted Boltzmann Machine (RBM) based initialization.

As we can see in the table, random initialization is worse than DNN with RBM pretraining by 0.9/0.6 in single GPU case. While in model averaging setup, random initialization

Nodes Data	1	2	4	8	16	32
SWB	14.7	–	–	15.1	15.1	15.2
CallHM	26.8	–	–	27.4	27.0	27.1
SWB	16.1	–	–	16.4	16.2	16.4
SWB2P3	21.0	–	–	21.8	21.7	21.7
SWB-Cell	27.4	–	–	27.3	27.4	27.8

**Table 1.** Comparison of WERs using different number of GPUs

	SWB		CallHome	
Nodes	1	32	1	32
random init	15.6	16.4	27.4	28.8
RBM pretraining	14.7	15.2	26.8	27.1

**Table 2.** Comparing RBM pretraining with random initialization

becomes even worse – 0.3/0.9 point more degradation on WER.

#### 5.4. Averaging frequency

Averaging frequency here is defined as the number of minibatch-SGD performed per model averaging. Due to the limitation of computing resource, we only did preliminary experiments on this. Minibatch size of 1024 is set as default, and we compare averaging frequency of 10 and 20. It is shown in Table 3 that an averaging frequency of 10 gives slight worse speedup but a better decoding WER. The tradeoff between lower averaging frequency (i.e. better speedup) and better training accuracy is within expectation in that frequent model averaging means steady gradient estimation.

#### 5.5. Minibatch Size

Table 4 compares two different minibatch size in model averaging setup.

#### 5.6. Learning Rate Schedule

Initial learning rate is increased in proportion to number of threads in model averaging setup. The reason for this is straight forward: Assume we have  $n$  minibatches of data for model training. When the model is trained using single thread, it gets updated  $n$  times. When data is distributed to  $m$

frequency	Speedup	SWB	CallHome
baseline	–	14.7	26.8
10	9.32	15.1	27.0
20	10.07	15.8	28.0

**Table 3.** Comparing different averaging frequencies

	SWB		CallHome		Speedup
nodes	1	16	1	16	
256	15.3	15.6	26.8	27.3	–
1024	14.7	15.1	26.8	27.0	9.32

**Table 4.** Comparing different minibatch size

	SWB		CallHome	
Nodes	1	16	1	16
Newbob	14.9	15.4	26.6	27.2
exponential	14.7	15.1	26.8	27.0

**Table 5.** Comparing learning rate schedule

machines, then each model gets updated  $n/m$  times. Since the effect of model averaging is mostly aggregating knowledge learnt from different data partition, the absolute change of model shall be compensated by  $m$  times.

We compare two learning rate schedules in this section. The first one is the default setup used in Kaldi nnet1 (Newbob). It starts with a initial learning rate of 0.32 and halves the rate when the improvement in frame accuracy on a cross-validation set between two successive epochs falls below 0.5%. The optimization terminates when the frame accuracy increases by less than 0.1%. Cross-validation is done on 10% of the utterances that are held out from the training data.

The second learning rate schedule is exponentially decaying. This method is used in [23, 11] and is shown to be superior to performance scheduling and power scheduling. In this work, it starts with the same initial learning rate as the first method (Newbob), and decrease to the final learning rate (which is set to be  $0.01 * \text{initial learning rate}$ ). The number of epochs is set to 15 in this task, which is set to be the same as Newbob scheduling.

As is shown in Table 5, these two learning rate scheduling methods give similar decoding results. However, exponential learning rate might need more tuning since it requires a initial learning rate, a final learning rate and predefined number of epochs to train.

#### 5.7. Online NG-SGD Matters

Table 6 compares plain SGD with NG-SGD in model averaging mode, and it shows NG-SGD is crucial to model training with parameter-averaging.

	SWB		CallHome	
Nodes	1	16	1	16
SGD	14.9	16.3	26.9	28.3
NG-SGD	14.7	15.1	26.8	27.0

**Table 6.** Comparing NG-SGD and naive SGD

## 6. CONCLUSION AND FUTURE WORK

In this work, we show that neural network training can be efficiently speeded up using model averaging. on a 300h Switchboard dataset, a 9.3x / 17x speedup could be achieved using 16 / 32 GPUs respectively, with limited decoding accuracy loss. We also show that model averaging benefits a lot from NG-SGD and RBM based pretraining. Preliminary experiments on minibatch size, averaging frequency and learning rate schedules are also presented.

Further accuracy improvement might be achieved if parallel training runs on top of serial training initialization. It would be interesting to see if sequence-discriminative training combines well with model averaging. Speedup factor could be further improved if CUDA aware MPI is used. Theory on convergence using model averaging is to be explored, which might be useful for guiding future development.

## 7. ACKNOWLEDGEMENTS

We would like to thank Karel Vesely and Daniel Povey who wrote the original "nnet1" neural network training code and natural gradient stochastic gradient descent upon which the work here is based. We would also like to thank Nelson Morgan, Forrest Iandola and Yuansi Chen for their helpful suggestions.

We acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported within this work (URL: <http://www.tacc.utexas.edu>).

## 8. REFERENCES

- [1] Stefano Scanzio, Sandro Cumani, Roberto Gemello, Franco Mana, and Pietro Laface, "Parallel implementation of artificial neural network training," in *Acoustics, Speech and Signal Processing (ICASSP), 2010 IEEE International Conference on*. IEEE, 2010, pp. 4902–4905.
- [2] Karel Veselý, Lukáš Burget, and František Grézl, "Parallel training of neural networks for speech recognition," in *Text, Speech and Dialogue*. Springer, 2010, pp. 439–446.
- [3] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu, "On parallelizability of stochastic gradient descent for speech dnns," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 235–239.
- [4] Frank Seide, Hao Fu, Jasha Droppo, Gang Li, and Dong Yu, "1-bit stochastic gradient descent and its application to data-parallel distributed training of speech dnns," in *Fifteenth Annual Conference of the International Speech Communication Association*, 2014.
- [5] Jeffrey Dean, Greg Corrado, Rajat Monga, Kai Chen, Matthieu Devin, Mark Mao, Andrew Senior, Paul Tucker, Ke Yang, Quoc V Le, et al., "Large scale distributed deep networks," in *Advances in Neural Information Processing Systems*, 2012, pp. 1223–1231.
- [6] Shanshan Zhang, Ce Zhang, Zhao You, Rong Zheng, and Bo Xu, "Asynchronous stochastic gradient descent for dnn training," in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6660–6663.
- [7] Xie Chen, Adam Eversole, Gang Li, Dong Yu, and Frank Seide, "Pipelined back-propagation for context-dependent deep neural networks," in *INTERSPEECH*, 2012.
- [8] Adam Coates, Brody Huval, Tao Wang, David Wu, Bryan Catanzaro, and Ng Andrew, "Deep learning with cots hpc systems," in *Proceedings of The 30th International Conference on Machine Learning*, 2013, pp. 1337–1345.
- [9] Xiaohui Zhang, Jan Trmal, Daniel Povey, and Sanjeev Khudanpur, "Improving deep neural network acoustic models using generalized maxout networks," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*. IEEE, 2014, pp. 215–219.
- [10] Yajie Miao, Hao Zhang, and Florian Metze, "Distributed learning of multilingual dnn feature extractors using gpus," 2014.
- [11] Daniel Povey, Xiaohui Zhang, and Sanjeev Khudanpur, "Parallel training of deep neural networks with natural gradient and parameter averaging," *arXiv preprint arXiv:1410.7455*, 2014.
- [12] D. Povey, A. Ghoshal, G. Boulianne, L. Burget, O. Glembek, N. Goel, M. Hannemann, P. Motlíček, Y. Qian, P. Schwartz, J. Silovský, G. Stemmer, and K. Veselý, "The kaldi speech recognition toolkit," in *ASRU*. IEEE, 2011.
- [13] Karel Veselý, Arnab Ghoshal, Lukáš Burget, and Daniel Povey, "Sequence-discriminative training of deep neural networks," in *INTERSPEECH*, 2013, pp. 2345–2349.
- [14] Geoffrey E Hinton, Simon Osindero, and Yee-Whye Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.
- [15] D Povey, D Kanevsky, B Kingsbury, B Ramabhadran, G Saon, and K Visweswariah, "Boosted mmi for feature and model space discriminative training," in *Proc. ICASSP*, 2008.

- [16] Frank Seide, Gang Li, and Dong Yu, “Conversational speech transcription using context-dependent deep neural networks,” in *Interspeech*, 2011, pp. 437–440.
- [17] Ryan McDonald, Keith Hall, and Gideon Mann, “Distributed training strategies for the structured perceptron,” in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 456–464.
- [18] Ryan McDonald, Mehryar Mohri, Nathan Silberman, Dan Walker, and Gideon S Mann, “Efficient large-scale distributed training of conditional maximum entropy models,” in *Advances in Neural Information Processing Systems*, 2009, pp. 1231–1239.
- [19] Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio, “Why does unsupervised pre-training help deep learning?,” *The Journal of Machine Learning Research*, vol. 11, pp. 625–660, 2010.
- [20] Noboru Murata and Shun-ichi Amari, “Statistical analysis of learning dynamics,” *Signal Processing*, vol. 74, no. 1, pp. 3–28, 1999.
- [21] Nicolas L Roux, Pierre-Antoine Manzagol, and Yoshua Bengio, “Topmoumoute online natural gradient algorithm,” in *Advances in neural information processing systems*, 2008, pp. 849–856.
- [22] Mark JF Gales, *The generation and use of regression class trees for MLLR adaptation*, University of Cambridge, Department of Engineering, 1996.
- [23] Alan Senior, Georg Heigold, Marc’Aurelio Ranzato, and Ke Yang, “An empirical study of learning rates in deep neural networks for speech recognition,” in *Acoustics, Speech and Signal Processing (ICASSP), 2013 IEEE International Conference on*. IEEE, 2013, pp. 6724–6728.